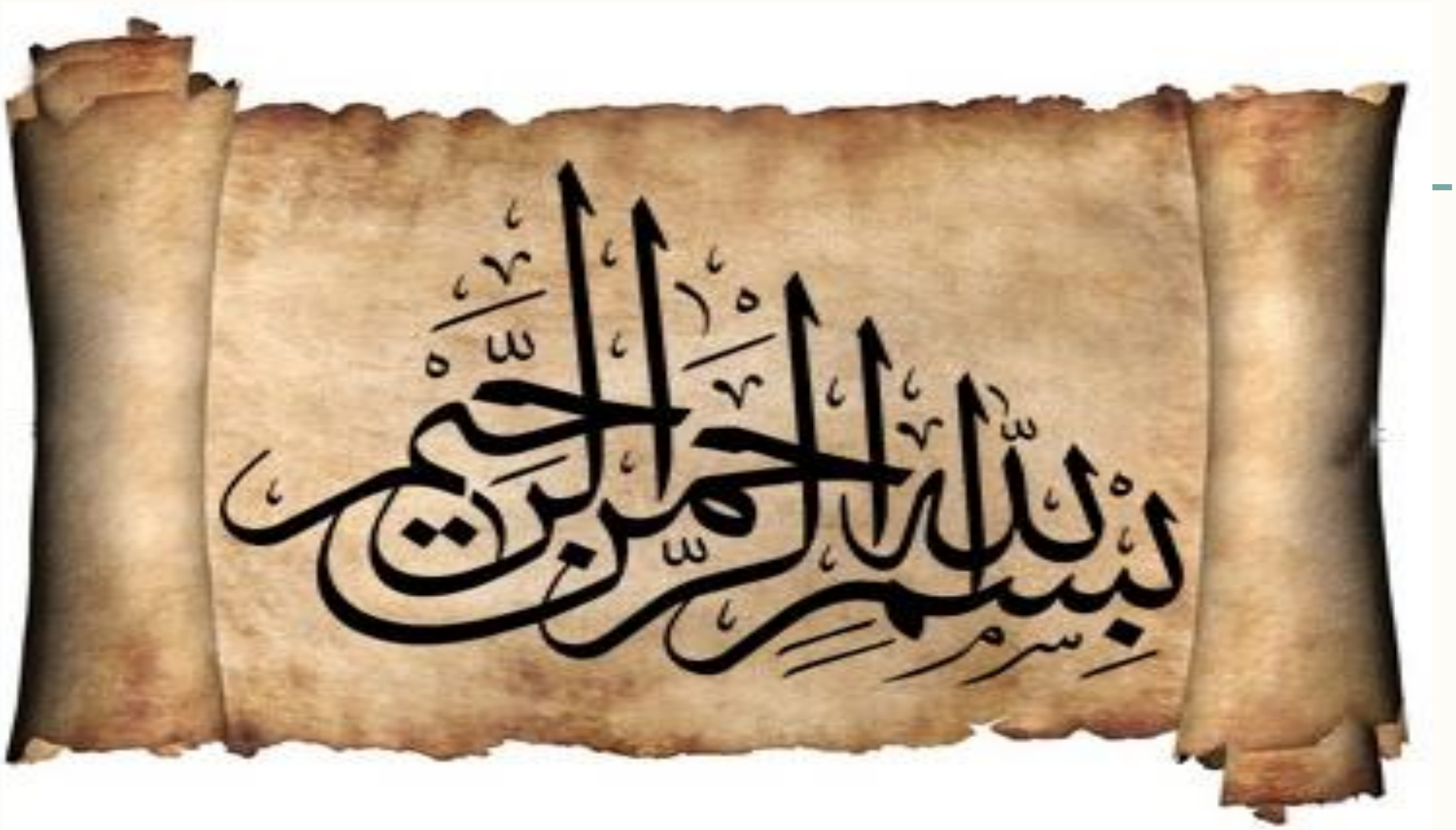

بسم الله الرحمن الرحيم

انتخاب متغیر در مدل های رگرسیونی

درس مدل خطی

زهره کمالی

۹۸-۹۹

Variable _selection

- مسئله انتخاب متغیر
- اهمیت انتخاب متغیر مناسب
- معیارهای انتخاب متغیر
- روش های کوچک سازی
- رگرسیون مرزی
- لاسو
- الاستیک نت
- منابع

انتخاب متغیر

variable selection

زیر مجموعه ای از متغیرهای رگرسیونی که باید در مدل به کار گرفته شوند ضرورتا بایستی تعیین شوند.

پیدا کردن چنین زیر مجموعه مناسب مدل از متغیرهای رگرسیونی مسئله انتخاب متغیر نامیده می شود.

ساختن یک مدل رگرسیونی که تنها شامل زیرمجموعه از متغیرهای رگرسیونی است دو واقعیت ناسازگار را دربردارد:

1) علاقه مندیم که مدل هرچه ممکن است متغیرهای رگرسیونی را دربرداشته باشد به نحوی که اطلاعات موجود در این عوامل بتوانند در پیش بینی مقدار Y اثرگذار باشد.

2) می خواهیم مدل در حد امکان دارای متغیرهای رگرسیونی کمتری باشد زیرا با افزایش تعداد متغیرهای رگرسیونی واریانس پیش بینی $Y^{\wedge}$ افزایش می یابد.همچنین هرچه متغیر رگرسیونی بیشتری در مدل باشد هزینه جمع آوری اطلاعات و ابقاء مدل افزایش می یابد.

اهمیت انتخاب متغیر
All models are wrong but some are useful

چرا ما نمی خواهیم همه ی متغیرهای توضیحی در مدل باشند؟؟

1. مدل واقعی معمولا پیچیده است.
2. مدل کامل و درست ، مجهول و ناشناخته است.
3. وجود متغیر اضافی در مدل باعث بیش برآورد می شود.

مدل های با تعداد متغیر کم همواره دلخواه تحلیل گران آماری بوده است، زیرا:

1. روابط بین متغیرهای علمی را به صورت ساده بیان کرده و قابلیت تفسیر ساده تری دارد.
2. هزینه های محاسباتی و صرف زمان زیاد روش های انتخاب متغیر را با چالش بسیارجدی در زمینه استفتده از داده با بعد بالا روبرو کرده است.
3. انتخاب یک مدل آماری خوب، مدلی است که استفاده از آن آسان، قابل فهم و قابل توضیح برای دیگران باشد.

معیار های انتخاب متغیر

- آزمون فرض:

این معیار ها زمانی قابل استفاده هستند که همه ی زیر مدل های ممکن شناخته شده باشند.

- روش های انقباضی:

برخی از معیار ها فقط زمانی کاربرد دارند که مدل مورد نظر برحسب پارامترها خطی باشند.

- روش های بیزی:

محاسبه ی این معیار ها ،برای تمام زیر مدل های ممکن ، محاسبات زیادی دارد به خصوص اگر متغیر های تصادفی زیاد باشد.

معیارهای انتخاب متغیر

- معیار هایی که نیکویی برازش را کنترل می کنند.
- معیار هایی که کیفیت پیش بینی مدل را ارزیابی می کنند.
- معیار هایی که به تفاوت بین توزیع واقعی و توزیع تخمین زده شده می پردازند.
- معیارهایی که احتمال پسین را تخمین می زنند.

همه ی مدل ها باید تا جایی که ممکن است ساده شوند،
ولی نه ساده تر از حد نیاز

معیارهایی که نیکویی برازش را کنترل می کنند:

R^2_{adj}

$R^2 = \frac{SSR}{SST}$

$R^2_{adj} = 1 - \frac{n-1}{n-p}(1-R^2)$

$MSE = \frac{SSE}{n-p}$

مدلی که MSE کمتری داشته باشد، بهتر است

VIF

$VIF_i = Cjj = \frac{1}{1-R_i^2}$

معیار هایی که کیفیت پیش بینی مدل را ارزیابی می کنند:

Mallow Cp

$Cp = \frac{1}{n}(RSS + 2d\sigma^2)$

CV & GCV

آمارهPRESS

PRESS= $\sum_{i=1}^{n}(Yi - \hat{Y}_i)^2$

معیار هایی که میزان تفاوت بین توزیع واقعی و توزیع تخمین زده شده می پردازند:

$$BIC=\frac{1}{n}(RSS+\log(n)d\sigma^2)$$

$$AIC=\frac{1}{n\sigma 2}(RSS+2d\sigma^2)$$

استفاده از آزمون فرض

1. **انتخاب پیشرو**

مزایای انتخاب پیشرو:این روش برای مسائلی که درابتدا نیاز به جمع آوری اطلاعات دارند،مناسب هستند

روش انجام این متد کاملا قابل فهم است.

معایب انتخاب پیشرو: متغیر وارد شده به مدل توسط این روش، قابل حذف نیست.

2. **حذف پسرو**

مزایای حذف پسرو: این روش، برای مسائلی که در آغاز فرضی بنا شده وسپس در جهت اثبات آن باشیم،مفید است.

روش انجام این متد کاملا قابل فهم است.

معایب حذف پسرو: متغیر حذف شده از مدل توسط این روش، قابل برگشت نیست.

روش های کوچک سازی
Shrinkage Methods

- وقتی p>n باشد، پارامترها نمی توانند با استفاده از براورد کمترین مربعات خطا (ols) به صورت یکتا تعریف شوند.
- در مواردی که متغیرها وابستگی دارند، براوردگر ols کارایی لازم را ندارد زیرا باعث بزرگی واریانس می شوند .
- مسئله هم خطی و کاربردی نبودن برآوردگر کمترین توان دوم خطا، منجر به توسعه برآوردگرهای اریب مانند

برآوردگرهای انقباضی (ریج و لاسو) می شود.

انقباض :

این روش شامل قرار دادن یک مدل شامل همه پیش بینی کننده های P پارامتر است.

با این حال ، ضرایب برآورد شده نسبت به تخمین حداقل مربعات به سمت صفر کاهش می یابد .

این کوچک شدن (همچنین به عنوان منظم سازی شناخته می شود) باعث کاهش واریانس می شود.

بستگی به نوع انقباضی که انجام می شود ، برخی از ضرایب دقیقا صفر تخمین زده می شوند از این رو ، روش های کوچک شدن نیز می توانند انتخاب متغیر انجام دهند.

رگرسیون مرزی(ریج) ولاسو

Ridge Regression & The Lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{6.8}$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s, \tag{6.9}$$

مقایسه روش رگرسیون مرزی و لاسو

رگرسیون مرزی برآوردگر اریب تولید می کند اما واریانس را کاهش می دهد.

رگرسیون مرزی همه ضرایب را به یک مقدار غیر صفر کاهش می دهد اما در روش لاسو مقداری از ضرایب دقیقا به صفر کاهش می یابد.

رگرسیون مرزی

Ridge Regression

$$\hat{\beta}_{Ridge} = \arg\min\left\{\|y - x\beta\|^2 + \lambda\|\beta\|^2\right\}$$

بنابراین رگرسیون مرزی ، محدودیت هایی را روی پارامترهای بتا در مدل خطی به وجود می آورد.
در این مورد کاری که ما انجام می دهیم این است که به جای مینیمم کردن مجموع مربعات خطا ما یک عبارت تاوان روی بتا ها نیز داریم این عبارت تاوان لاندا برابر مربع نرم بردار بتا است. این به این معنی است که اگربتاها مقادیر بزرگی اختیار کنند تابع بهینه تاوانیده می شود. ما ترجیح می دهیم که بتاها مقادیرکوچک بگیرند یا این که به صفر نزدیک شوندکه عبارت تاوان کوچک شود.

رگرسیون مرزی

Ridge Regression

λ پارامتر کنترل است.

λ معمولاً به گونه ای انتخاب می شود که برآوردگر رگرسیون مرزی ، میانگین مربعات خطای کمتری نسبت به برآوردگر حداقل مربعات داشته باشد.

این برآوردگر در مقایسه با برآوردگر کمترین مربعات خطا اریب است اما واریانس کمتری دارد.

این براوردگر معمولا در مواردی با تعداد دامنه های بالا (تعداد متغیرهای مستقل زیاد n>p) کاربرد دارد.

- ایرادات وارده به روش رگرسیون مرزی این است که :

این برآوردگرها تا حدودی نیزاریب هستند و مانع از صفر شدن برآورد ضرائب رگرسیونی می شوند.

لاسو

The Lasso

این روش شبیه به روش رگرسیون مرزی است با این تفاوت که به جای استفاده از تابع تاوان درجه دو از تابع تاوان قدر مطلق ضرایب استفاده می کند و باعث می شود که بعضی از ضرایب دقیقا به صفر کاهش یابند.

$$\beta_{Lasso} = \arg\min \left\{ \|y - x\,\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

مزایا و معایب لاسو

مزایای روش لاسو:

روش لاسو به لحاظ پایداری و دقت پیش بینی برآوردگرها عملکرد قابل قبولی از خود نشان داده است.

نسبت به رگرسیون مرزی اکثر مواقع دقت پیش بینی بالایی دارد.

معایب روش لاسو:

1. محاسبات بسیار پیچیده.
2. فاقد ویژگی نااریبی است.
3. یک مقدار تاوان ثابت برای ضرایب در نظر میگیرد.
4. عدم ثبات کافی در انتخاب متغیرهای موثر زمانی که داده ها شامل گروه هایی از متغیرهای پیش بینی به شدت به هم وابسته هستند.
5. مشکل دیگر لاسو درارتباط با مسائلی است که تعداد متغیرهای توضیحی بزرگتر از حجم نمونه است.

روش الاستیک نت:

برآورد ضرائب در این روش به فرم زیر است:

$$\hat{\beta}_{NEN} = \arg\min\left\{\|y - x\,\beta\|^2\right\} + \lambda\sum_{j=1}^{p}|\beta_j| + \lambda(1-\alpha)\sum_{j=1}^{p}\beta_j{}^2$$

ناحیه ی تاوان در این روش ترکیبی محدب از ناحیه تاوان در روش های لاسو و رگرسیون مرزی است.
این فرم از تابع تاوان علاوه بر توانایی صفر براورد کردن برخی از ضرایب، توانایی برابر براورد کردن ضرایب متغیرهایی که اثر یکسانی روی متغیر پاسخ دارند یا بشدت همبسته هستند نیز دارد.

مزایای روش فوق:

1. این روش هم از نظر دقت پیش بینی و هم از نظر انتخاب مدل صحیح از دیگر روش های مذکور بهتر عمل می کند.
2. این روش از نظر تمایز بین متغیرهای مؤثر وغیرمؤثر همبسته نیز از دیگر روش های مذکور بهتر عمل می کند.

منابع

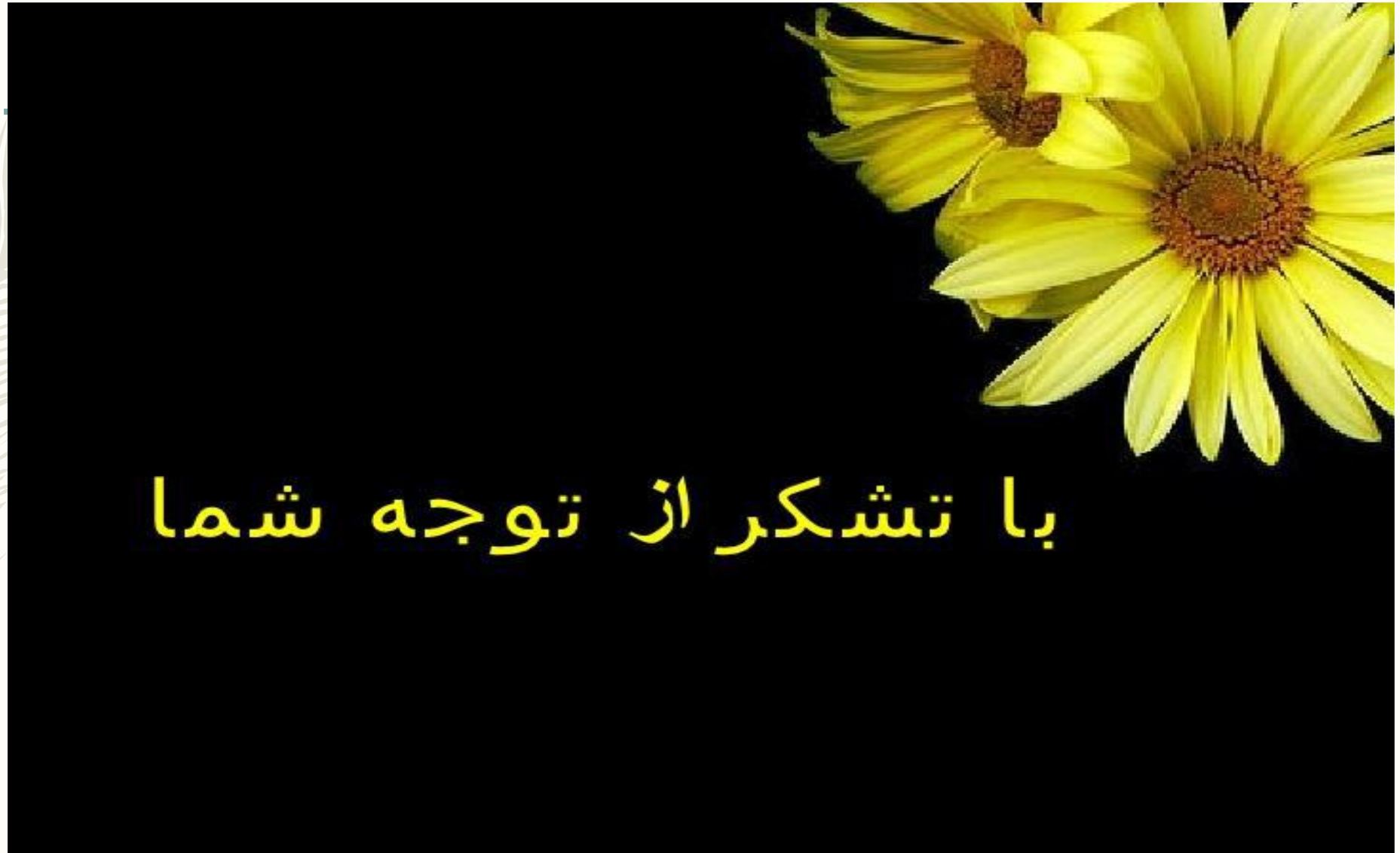
مقدمه ای بر تحلیل رگرسیون خطی (داگلاس مونتگمری،الیزابت پک)

The elements of statistical learning ((Springer Series in Statistics) Trevor Hastie, Robert Tibshirani, Jerome Friedman)

An introduction to statistical learning (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani)

Joseph B.kadane; Nicole A . Lazar; Methods and Criteria for model selection

Alvin C. Rencher and G. Bruce Schaalje ; LINEAR MODELS IN STATISTICS


با تشکر از توجه شما